Higher-order Comparisons of Sentence Encoder Representations

vqc439, vqc439; Kulmizev, Artur ; Hill, Felix ; Low, Daniel M. Low; Søgaard, Anders

Published in:
Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Higher-order Comparisons of Sentence Encoder Representations

Mostafa Abdou[†] [*] **Artur Kulmizev**[♣] **Felix Hill**[♢] **Daniel M. Low**[♠] **Anders Søgaard**[†]
[†]Department of Computer Science, University of Copenhagen
[♣]Department of Linguistics and Philology, Uppsala University
[♢]DeepMind [♠]Program in Speech and Hearing Bioscience and Technology,
Harvard Medical School-MIT

Abstract

Representational Similarity Analysis (RSA) is a technique developed by neuroscientists for comparing activity patterns of different measurement modalities (e.g., fMRI, electrophysiology, behavior). As a framework, RSA has several advantages over existing approaches to interpretation of language encoders based on probing or diagnostic classification: namely, it does not require large training samples, is not prone to overfitting, and it enables a more transparent comparison between the representational geometries of different models and modalities. We demonstrate the utility of RSA by establishing a previously unknown correspondence between widely-employed pretrained language encoders and human processing difficulty via eye-tracking data, showcasing its potential in the interpretability toolbox for neural models.

1 Introduction

Examining the parallels between human and machine learning is a natural way for us to better understand the former and track our progress in the latter. The "black box" aspect of neural networks has recently inspired a large body of work related to interpretability, i.e. understanding of representations that such models learn. In NLP, this push has been largely motivated by linguistic questions, such as: *what linguistic properties are captured by neural networks?* and *to what extent do decisions made by neural models reflect established linguistic theories?* Given the relative recency of such questions, much work in the domain so far has been focused on the context of models in isolation (e.g. *what does model* **X** *learn about linguistic phenomenon* **Y**?) In order to more broadly understand models' representational tendencies, however, it is vital that such questions be formed not only with other models in mind, but also other representational methods and modalities (e.g. behavioral data, fMRI measurements, etc.). In context of the latter concern, the present-day interpretability toolkit has not yet been able to afford a practical way of reconciling this.

In this work, we employ Representational Similarity Analysis (RSA) as a simple method of interpreting neural models' representational spaces as they relate to other models and modalities. In particular, we conduct an experiment wherein we investigate the correspondence between human processing difficulty (as reflected by gaze fixation measurements) and the representations induced by popular pretrained language models. In our experiments, we hypothesize that there exists an overlap between the sentences which are difficult for humans to process and those for which per-layer encoder representations are least correlated.

Our intuition is that such sentences may exhibit factors such as low-frequency vocabulary, lexical ambiguity, and syntactic complexity (e.g. multiple embedded clauses), etc. that are uncommon in both standard language and, relatedly, the corpora employed in training large-scale language models. In the case of a human reader, encountering such a sentence may result in a number of processing delays, e.g. longer aggregate gaze duration. In the case of a sentence encoder, an uncommon sentence may lead to a degradation of representations in the encoder's layers, wherein a lower layer might learn to encode vastly different information than a higher one. Similarly, different models' representations may emphasize different aspects of these more complex sentences and therefore diverge from each other. With this in mind, our hypothesis is that sentences which are difficult for humans to process are likely to have divergent representations within models' internal layers and between different models' layers.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5838–5845,
Hong Kong, China, November 3–7, 2019.

Understanding and analysing language encoders In recent years, some prominent efforts towards interpreting neural networks for NLP have included: developing suites that evaluate network representations through performance on downstream tasks (Conneau et al., 2017a; Wang et al., 2018; McCann et al., 2018); analyzing network predictions on carefully curated datasets (Linzen et al., 2016; Marvin and Linzen, 2018; Gulordava et al., 2018; Loula et al., 2018; Dasgupta et al., 2018; Tenney et al., 2018); and employing diagnostic classifiers to assess whether certain classes of information are encoded in a model's (intermediate) representations (Adi et al., 2016; Chrupała et al., 2017; Hupkes et al., 2017; Belinkov et al., 2017).

While these approaches provide valuable insights into how neural networks process a large variety of phenomena, they rely on decoding accuracy as a probe for encoded linguistic information. If properly biased, this means that they can detect whether information is encoded in a representation or not. However, they do not allow for a direct comparison of representational structure between models. Consider a toy dataset of five sentences of interest and three encodings derived from quite different processing models; a hidden state of a trained neural language model, a *tf-idf* weighted bag-of-words representation, and measurements of fixation duration from an eye-tracking device. Probing methods do not allow us to quantify or visualise, for each of these encoding strategies, how the encoder's responses to the five sentences relate to each other. Moreover, probing methods would not directly reveal whether the fixations from the eye-tracking device aligned more closely with the tf-idf representation or the states of the neural language model. In short, while probing classifier methods can establish if phenomena are separable based on the provided representations, they do not tell us about the overall geometry of the representational spaces. RSA, on the other hand, provides a basis for higher-order comparisons between spaces of representations, and a way to visualise and quantify the extent to which they are isomorphic.

Indeed, RSA has seen a modest introduction within interpretable NLP in recent years. For example, Chrupała et al. (2017) employed RSA as a means of correlating encoder representations of speech, text, and images in a post-hoc analysis of a multi-task neural pipeline. Similarly, Bouchacourt and Baroni (2018) used the framework to measure the similarity between input image embeddings and the representations of the same image by an agent in an language game setting. More recently, Chrupała and Alishahi (2019) correlated activation patterns of sentence encoders with symbolic representations, such as syntax trees. Lastly, similar to our work here, Abnar et al. (2019) proposed an extension to RSA that enables the comparison of a single model in the face of isolated, changing parameters, and employed this metric along with RSA to correlate NLP models' and human brains' respective representations of language. We hope to position our work among this brief survey and further demonstrate the flexibility of RSA across several levels of abstraction.

2 Representational Similarity Analysis

RSA was proposed by Kriegeskorte et al. (2008) as a method of relating the different representational modalities employed in neuroscientific studies. Due to the lack of correspondence between the activity patterns of disparate measurement modalities (e.g. brain activity via fMRI, behavioural responses), RSA aims to abstract away from the activity patterns themselves and instead compute representational dissimilarity matrices (RDMs), which characterize the information carried by a given representation method through dissimilarity structure.

Given a set of representational methods (e.g., pretrained encoders) M and a set of experimental conditions (sentences) N, we can construct RDMs for each method in M. Each cell in an RDM corresponds to the dissimilarity between the activity patterns associated with pairs of experimental conditions $n_i, n_j \in N$, say, a pair of sentences. When $n_i = n_j$, the dissimilarity between an experimental condition and itself is intuitively 0, thus making the $N \times N$ RDM symmetric along a diagonal of zeros (Kriegeskorte et al., 2008).

The RDMs of the different representational methods in M can then be directly compared in a Representational *Similarity* Matrix (RSM). This comparison of RDMs is known as second-order analysis, which is broadly based on the idea of a *second-order isomorphism* (Shepard and Chipman, 1970). In such an analysis, the principal point of comparison is the match between the dissimilarity structure of the different representa-

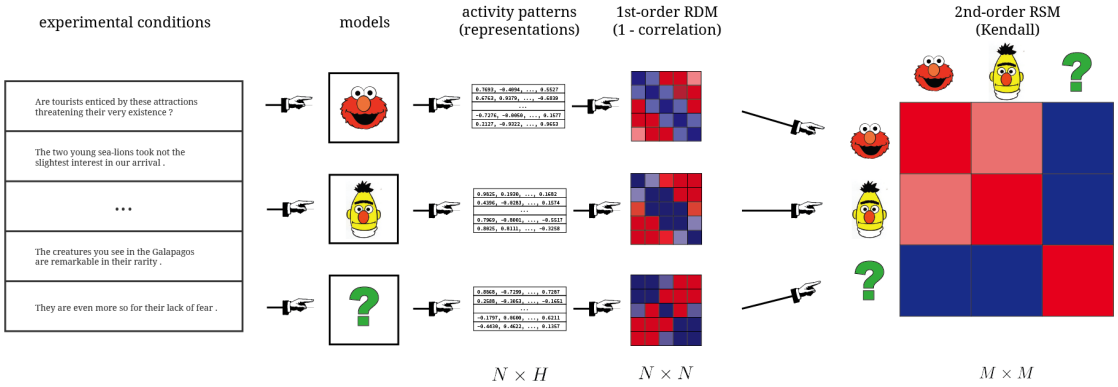

Figure 1: An example of first- and second-order analyses, where $N = \#$ of experimental conditions, $M = \#$ of models, and $H = \#$ of activity patterns observed for a given model (i.e. dimensionality). The right-most side of the figure depicts a representational similairty matrix (RSM) of correlations between RDMs.

tional methods. Intuitively, this can be expressed through the notion of *distance between distances*, and is thus related to Earth Mover's Distance (Rubner et al., 2000).[1] Figure 1 shows an illustration of the first and second order analyses for pretrained language encoders.

Note that RSA is meaningfully different from, and complementary to, methods that employ saturating functions of representation distances (e.g. decoding accuracy, mutual information), which suffer from **(a)** a ceiling effect: being able to distinguish experimental phenomenon A from B with with an accuracy of 100% and experimental phenomenon C from D with an accuracy of 100% does not mean that the distance between A and B is the same as that between C and D; and **(b)** discretization (Nili et al., 2014).

We follow Kriegeskorte et al. (2008) in using the correlation distance of experimental condition pairs $n_i, n_j \in N$ as a dissimilarity measure, where $\bar{n_i}$ is the mean of n_i's elements, $\cdot$ is the dot product, and $\|$ is the l_2 norm: $corr(x) = 1 - \frac{(n_i - \bar{n_i}) \cdot (n_j - \bar{n_j})}{\|(n_i - \bar{n_i}\|_2 \|(n_j - \bar{n_j}\|_2}$. Compared to other measures, correlation distance is preferable as it normalizes both the mean and variance of activity patterns over experimental conditions. Other popular measures include the Euclidean distance and the Malahanobis distance (Kriegeskorte et al., 2006).

3 Fixation Duration and Encoder Disagreement

Gaze fixation patterns have been shown to strongly reflect the online cognitive processing demands of human readers (Raney et al., 2014; Ashby et al., 2005) and to be dependent upon a number of linguistic factors (Van Gompel, 2007). Specifically, it has been demonstrated that word frequency, syntactic complexity, and lexical ambiguity play a strong part in determining which sentences are difficult for humans to process (Rayner and Duffy, 1986; Duffy et al., 1988; Levy, 2008).

Using the RSA framework, we aim to explore how gaze fixation patterns and the linguistic factors associated with sentence processing difficulty relate to the representational spaces of popular language encoders. Namely, we hypothesize that, for a given sentence, disagreement between hidden layers corresponds to processing difficulty. Because layer disagreement for a sentence measures the extent to which two layers (e.g. within BERT) disagree with each other about the pairwise similarity of the sentence (with other sentences in the corpus), a sentence with high layer disagreement will have unstable similarity relationships to other sentences in the corpus. This indicates that it has a degraded encoder representation. Going further, we also hypothesize that models' representations of said sentences may be confounded, in part, by factors that are known to influence humans.

Eye-tracking data For our experiments, we make use of the Dundee eye-tracking corpus (Kennedy et al., 2003), the English part of which consists of eye-movement data recorded as 10 native participants read 2,368 sentences from 20 newspaper articles. We consider the following fixation features: TOTAL FIXATION DURATION and FIRST PASS DURATION. For each of the features, we first take the average of the measurements recorded for all 10 participants per word, then ob-

[1] More precisely, our measure of dissimilarity between experimental conditions is analogous to *ground distance* and dissimilarity between RDMs to *earth mover's distance.*

tain sentence-level annotations by summing the measurements of all words in a sentence and dividing by its length. The result of this is two vectors V_{totfix} and $V_{firstpass}$ of length $2,368$, where each cell in the vector corresponds to a sentence's average total fixation and average first pass duration, respectively.

Syntactic complexity, word frequency, and lexical ambiguity We also consider the three following linguistic features which affect processing difficulty. For each of the following the result is also a vector of length $2,368$ where each cell corresponds to a sentence:

a. the average word log frequency per sentence extracted from the British National Corpus (Leech, 1992), $V_{logFreq.}$.
b. the average number of senses per word per sentence extracted from WordNet (Miller, 1995), $V_{wordSense}$.
c. Yngve scores, a standard measure of syntactic complexity based on cognitive load (Yngve, 1960) , V_{Yngve}.

Pretrained encoders We conduct our analysis on pretrained BERT-large (Devlin et al., 2018) and ELMo (Peters et al., 2018), two widely employed contextual sentence encoders. To obtain a representation of a sentence from a given layer L, we perform mean-pooling over the time-steps which correspond to the words of a sentence, obtaining a vector representation of the sentence. Mean-pooling is a common approach for obtaining vector representations of sentences for downstream tasks (Peters et al., 2018; Conneau et al., 2017b). We refer to ELMo's lowest layer as E1, BERT's 11th layer as B11, etc.

RDMs We construct an RDM (see §2) for each contextual encoder's layers. Each RDM is a $2,368 \times 2,368$ matrix which represents the dissimilarity structure of the layer, (i.e., each row vector in the matrix contains the dissimilarity of a given sentence to every other sentence). We then compute the correlations between the two different RDMs. For our evaluation of how well the representational geometry of a layer correlates to another, we employ Kendall's τ_A as suggested in Nili et al. (2014), computing the pairwise correlation for each two corresponding rows in two RDMs. This second-order analysis gives us a pairwise relational similarity vector $V_{Corr_{L_i-L_j}}$ of length $2,368$, which has the correlations between two layers L_i and L_j's RDMs for each of the sentences.

Third-order analysis The final part of our analysis involves computing correlations (Spearman's ρ) of $\{V_{Corr_{L_i-L_j}}, V_{logFreq}, V_{Yngve}, V_{wordSense}\}$ with each of V_{totfix} and $V_{firstpass}$. The results from this are shown in Table 1. The top section of the table shows correlations when L_i and L_j are the three final adjacent layers in BERT and ELMo. The middle section shows the results for top three BERT layer pairs L_i and L_j which maximize the correlation scores. The final section shows correlation with the linguistic features. Finally, Figure 2 shows Spearman's ρ correlations between $V_{Corr_{L_i-L_j}}$ and each of V_{totfix}, and V_{Yngve} for all combinations of the 24 BERT layers.

4 Discussion

Our results show highly significant negative correlations between $V_{Corr_{L_i-L_j}}$ and sentence gaze fixation times. These findings confirm the hypothesis that the sentences that are most challenging for humans to process, are the sentences **(a)** the layers of BERT disagree most on among themselves; and **(b)** that ELMo and BERT disagree most on, indicating that there may be common factors which affect human processing difficulty and result in disagreement between layers. By Layer disagreement we refer to the expression $1 - V_{Corr_{L_i-L_j}}$. It is important to note that these encoders are trained with a language modelling objective, unlike models where reading behaviour is explicitly modelled (Hahn and Keller, 2016) or predicted (Matthies and Søgaard, 2013). Indeed, the similarities here emerge naturally as a function of the task being performed. This can be seen as analogous to the case of similarities observed between neural networks trained to perform object recognition and spatio-temporal cortical dynamics (Cichy et al., 2016).

Syntactic complexity Figure 2 shows that, for all combinations of BERT layers, total fixation time and Yngve scores have strong negative and positive correlations (respectively) with layer disagreement. Furthermore, we observe that disagreement between middle layers seems to show the strongest correlation with Yngve scores. To confirm this, we split the correlations into four groups: "low" ($i, j \in [1, 8]$), "middle" ($i, j \in$

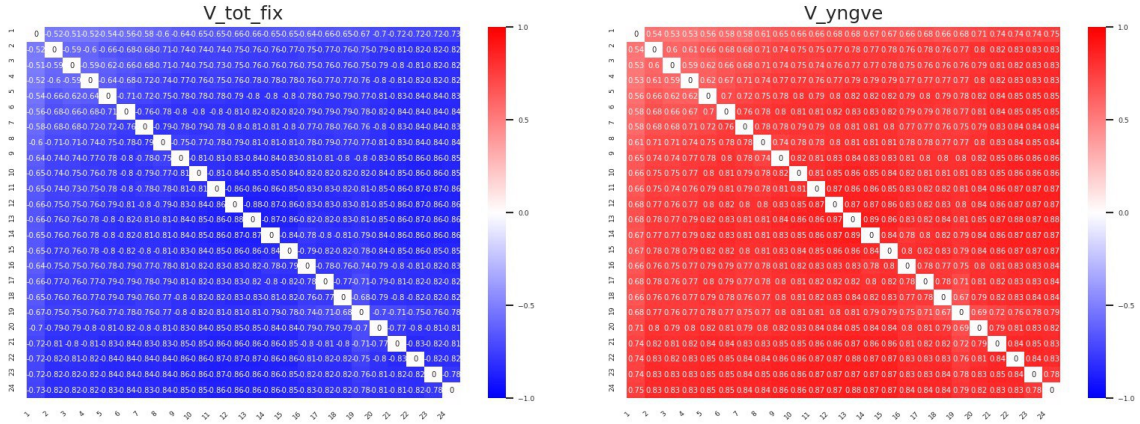

Figure 2: RSMs showing (Spearman's ρ) correlation between disagreement among layers i and j ($V_{Corr_{L_i-L_j}}$) and V_{totfix} (left) and V_{Yngve} (Right). BERT layers are denoted with numbers from 1 (topmost) to 24 (lowest).

Layer Disagreement	Total Fixation	First Pass Duration
E1-B22	-0.46	-0.46
E2-B23	-0.66	-0.67
E3-B24	-0.22	-0.23
B11-B12	-0.88	-0.87
B12-B13	-0.87	-0.85
B10-B21	-0.87	-0.86
Linguistic Features		
Log Freq.	-0.20	-0.19
Avg. Senses per Word	-0.007*	-0.004*
Yngve Score	0.66	0.66

Table 1: Spearman's ρ between $V_{Corr_{L_i-L_j}}$, $V_{logFreq.}$, $V_{wordSense}$, V_{Yngve} and each of V_{totfix} and $V_{firstpass}$. All correlations significant with $p < 0.0001$ after Bonferroni correction unless marked with *.

$[9, 16]$), "high" ($i, j \in [17, 24]$), and "out" ($|i - j| > 7$), with the latter representing out-of-group correlations (e.g. $Corr_{L_1-L_{24}}$). To account for correlations between disagreeing adjacent layers (e.g. $|i - j| = 1$) and Yngve scores being higher (as a possible confounding factor), we also distinguish layers as either "adjacent" or "non-adjacent". Considering these two factors as three- and two-leveled independent variables respectively, we conduct a two-way analysis of variance. The analysis reveals that the effect of group is significant at $F(3, 275) = 78.47, p < 0.0001$, with "low" ($\mu$ = 0.65, σ = 0.08), "middle" (μ = 0.84, σ = 0.03), "high" (μ = 0.80, σ = 0.05), and "out" (μ = 0.80, σ = 0.05). Neither the effect of adjacency nor its interaction with group proved to be significant.

This can be seen as (modest) support for the findings of previous work (Blevins et al., 2018; Tenney et al., 2019): namely, that the intermediate layers of neural language models encode the most syntax, and are therefore possibly more sensitive towards syntactic complexity. A very similar pattern is observed for total fixation time. When considered together with the correlation between V_{Yngve} and fixation times, this indicates a tripartite affinity between layer disagreement, syntactic complexity, and fixation.

Lexical Ambiguity and Word Frequency Finally, we observe that $V_{logFreq.}$ has a moderate correlation with both fixation time and layer disagreement and that $V_{wordSense}$ is nearly uncorrelated to both. Detailed plots of the latter can be found in Appendix A.

5 Conclusion

We presented a framework for analyzing neural network representations (RSA) that allowed us to relate human sentence processing data with language encoder representations. In experiments conducted on two widely used encoders, our findings show that sentences which are difficult for humans to process have more divergent representations both intra-encoder and between different encoders. Furthermore, we lend modest support to the intuition that a model's middle layers encode comparatively more syntax. Our framework offers insight that is complimentary to decoding or probing approaches, and is particularly useful to compare representations from across modalities.

Acknowledgements

We would like to thank Vinit Ravishankar, Matt Lamm, and the anonymous reviewers for their helpful comments. Mostafa Abdou and Anders Søgaard are supported by a Google Focused Research Award and a Facebook Research Award.

References

Abnar, S., Beinborn, L., Choenni, R., and Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Ashby, J., Rayner, K., and Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A*, 58(6):1065–1086.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1–10.

Blevins, T., Levy, O., and Zettlemoyer, L. (2018). Deep rnns encode soft hierarchical syntax. *arXiv preprint arXiv:1805.04218*.

Bouchacourt, D. and Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

Chrupała, G. and Alishahi, A. (2019). Correlating neural and symbolic representations of language. *arXiv preprint arXiv:1905.06401*.

Chrupała, G., Gelderloos, L., and Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *arXiv preprint arXiv:1702.01991*.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017a). Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017b). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., and Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duffy, S. A., Morris, R. K., and Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of memory and language*, 27(4):429–446.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hahn, M. and Keller, F. (2016). Modeling human reading with neural attention. *arXiv preprint arXiv:1608.05604*.

Hupkes, D., Veldhoen, S., and Zuidema, W. (2017). Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *arXiv preprint arXiv:1711.10203*.

Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Leech, G. N. (1992). 100 million words of english: the british national corpus (bnc).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *arXiv preprint arXiv:1611.01368*.

Loula, J., Baroni, M., and Lake, B. M. (2018). Rearranging the familiar: Testing compositional generalization in recurrent networks. *arXiv preprint arXiv:1807.07545*.

Marvin, R. and Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Matthies, F. and Søgaard, A. (2013). With blinkers on: Robust prediction of eye movements across readers. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 803–807.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Raney, G. E., Campbell, S. J., and Bovee, J. C. (2014). Using eye movements to evaluate the cognitive processes involved in text comprehension. *Journal of visualized experiments: JoVE*, (83).

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Rubner, Y., Tomasi, C., and Guibas, L. (2000). The earth mover's distance as a metric for image retrieval. In *IJCV*.

Shepard, R. N. and Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., et al. (2018). What do you learn from context? probing for sentence structure in contextualized word representations.

Van Gompel, R. P. (2007). *Eye movements: A window on mind and brain*. Elsevier.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

A Correlation Heatmaps

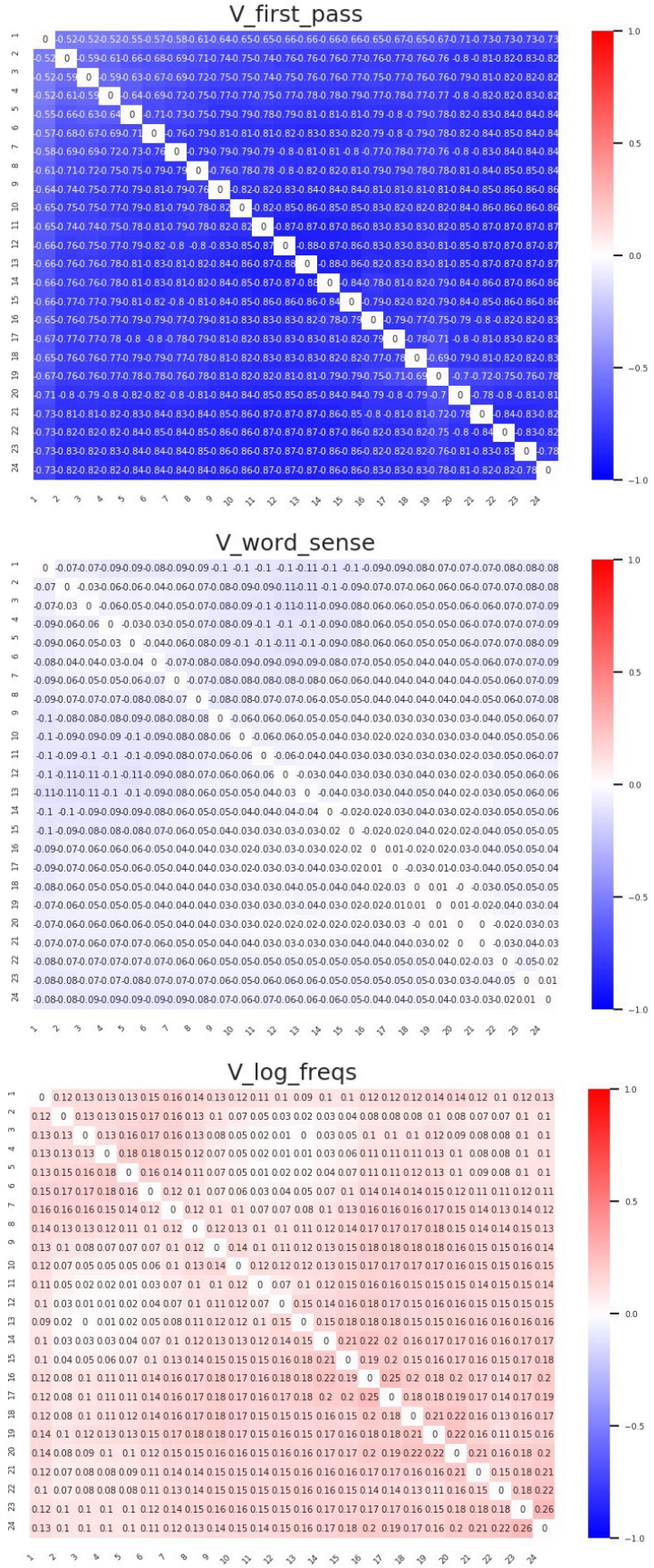

Figure 3: RSM heatmaps showing (Spearman's ρ) correlation between disagreement among layers i and j ($V_{Corr_{L_i - L_j}}$) and **(a)** $V_{firstpass}$ (top), **(b)** $V_{wordSense}$ (middle) and, **(c)** $V_{logFreq}$ (bottom).